# DISTRIBUTIVE SIMILARITY COMPOSITION FOR CLIP-BASED CLASS-INCREMENTAL LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Parameter-efficient fine-tuning (PEFT) provides an effective way to adapt vision-language models to sequential tasks while preserving their pre-trained backbones. However, when separate task-specific modules are maintained, task-agnostic inference in class-incremental learning typically relies on routing or retrieval to activate the appropriate module, making final predictions vulnerable to selection errors. In this paper, we exploit a distinctive property of CLIP to replace selection-based inference with task-representation composition. Since CLIP performs classification through inner-product similarities, and the inner product distributes over vector addition, the similarity of a composed visual representation can be decomposed into the additive contributions of individual task-wise representations. Based on this observation, we propose Distributive Similarity Composition for Incremental Learning (DISCIL), a three-stage rehearsal-free framework. Specifically, DISCIL freezes the CLIP textual space and trains a visual LoRA for each incoming task. Motivated by the observation that visual features are more crowded than their corresponding text anchors, each LoRA is optimized to produce high responses to relevant textual anchors while suppressing unrelated responses. To improve inference efficiency and stability, we further introduce an incremental two-teacher consolidation strategy that distills the previous merged LoRA and the current task LoRA into a new merged LoRA using mean squared error and cosine objectives. Consequently, the final merged LoRA requires only one visual forward pass, and its output is classified using all seen-class text anchors together with an auxiliary cumulative cosine classifier. Extensive experiments on multiple benchmarks demonstrate that DISCIL achieves strong continual recognition performance with favorable parameter and inference efficiency.

## 1 INTRODUCTION

Class-incremental learning (CIL) aims to continuously learn new visual categories while retaining the ability to recognize previously learned ones (Wu et al., 2019; Wang et al., 2023b). Large-scale vision-language models, particularly CLIP (Radford et al., 2021), provide a promising foundation for this problem through their transferable visual and textual representations (Luo et al., 2025; Garg et al., 2024; Li et al., 2026; Liu et al., 2025). However, fully fine-tuning such models is computationally expensive and may distort their pre-trained knowledge, motivating the use of parameter-efficient fine-tuning (PEFT) in continual learning (Kurniawan et al., 2024; Huang et al., 2024b).

A common PEFT strategy is to assign lightweight task-specific parameters, such as Prompts (Li & Liang, 2021), Adapters (He et al., 2021), and LoRA (Hu et al., 2022), while freezing the pre-trained backbone. Prompt-based methods, including L2P (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), and CODA-Prompt (Smith et al., 2023), have demonstrated the effectiveness of this strategy in rehearsal-free CIL. However, maintaining multiple PEFT modules creates a fundamental problem during task-agnostic inference: the model must determine which module should be activated for an input with an unknown task identity. As illustrated in Fig. 1(a), existing methods commonly rely on prompt retrieval, query–key matching, or expert routing. These selection-oriented mechanisms introduce additional inference steps and make the final prediction vulnerable to module-selection errors (Kim et al., 2024; Chen et al., 2025; Yu et al., 2024). This motivates a different question: **can task-specific PEFT modules be jointly utilized without selecting a task-relevant module?**



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Toward this question, we find that the property of CLIP's similarity-based classifier provides a natural solution. Unlike conventional visual classifiers with a learned classification head, CLIP represents each class $c$ using a textual anchor $\mathbf{t}_c$ and predicts the class with the highest image–text similarity:

$$\hat{y} = \arg\max_c \hat{\mathbf{z}}^\top \hat{\mathbf{t}}_c, \tag{1}$$

where $\hat{\mathbf{z}}$ and $\hat{\mathbf{t}}_c$ denote normalized visual and textual features, respectively. Thus, *visual representations produced by different task-specific modules can be interpreted within the same textual-anchor space.*

Supposing that each task-specific module produces an adapted visual representation $\hat{\mathbf{z}}^{(k)}$, we compose these task-wise representations as

$$\widetilde{\mathbf{z}} = \sum_{k=1}^{K} \alpha_k \hat{\mathbf{z}}^{(k)}, \qquad \hat{\mathbf{z}} = \frac{\widetilde{\mathbf{z}}}{\|\widetilde{\mathbf{z}}\|_2}. \tag{2}$$

For an arbitrary anchor $\hat{\mathbf{t}}_c$, the resulting similarity can be decomposed as

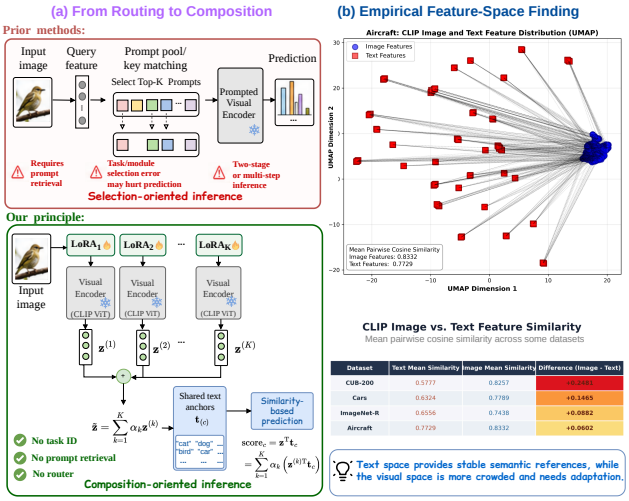


Figure 1: (a) Existing task-specific PEFT methods typically rely on prompt retrieval or expert routing to select a suitable module. In contrast, we directly compose task-wise visual representations and evaluate them using CLIP's shared textual anchors. (b) UMAP visualization on Aircraft and mean pairwise cosine similarities on four datasets show that visual features are substantially more crowded than their corresponding textual anchors. This motivates preserving the stable textual space while adapting the visual encoder.

$$\hat{\mathbf{z}}^\top \hat{\mathbf{t}}_c = \frac{1}{\|\widetilde{\mathbf{z}}\|_2} \left( \sum_{k=1}^{K} \alpha_k \hat{\mathbf{z}}^{(k)} \right)^\top \hat{\mathbf{t}}_c = \frac{1}{\|\widetilde{\mathbf{z}}\|_2} \sum_{k=1}^{K} \alpha_k \left( \hat{\mathbf{z}}^{(k)\top} \hat{\mathbf{t}}_c \right). \tag{3}$$

The normalization term is independent of the class and therefore does not affect class ranking. Equation 3 shows that the similarity of a composed representation is equivalent to accumulating the contributions of individual task-wise representations to the same textual anchor. Therefore, task-specific modules can be treated as complementary visual representations rather than competing candidates that must be selected, as illustrated in Fig. 1(a).

Based on this insight, we propose **Di**stributive **S**imilarity **C**omposition for **I**ncremental **L**earning (DISCIL), a rehearsal-free framework that replaces task-module routing with additive representation composition. Although the distributive property always holds mathematically, effective composition requires each task representation to produce selective responses.

Specifically, for an image $\mathbf{x}$ from task $k$ with ground-truth label $y \in \mathcal{C}_k$, a composition-compatible representation should produce a high response to the relevant textual anchor and low responses to unrelated textual anchors:

$$\hat{\mathbf{z}}^{(k)}(\mathbf{x})^\top \hat{\mathbf{t}}_c \text{ is high for } c = y, \quad \text{and low for } c \in \mathcal{C}_{1:k} \setminus \{y\}. \tag{4}$$

Such selective responses allow the relevant task representation to contribute more strongly to the correct class while reducing interference from unrelated classes during representation composition.

Our empirical analysis further indicates *where continual adaptation should be performed*. As shown in Fig. 1(b), visual class features exhibit consistently higher pairwise similarity than their corresponding textual anchors, indicating a more crowded and less discriminative visual space. In contrast, CLIP's text encoder has been trained on highly diverse image–text pairs and has already learned relatively stable semantic representations for class names. Freezing the pre-trained text encoder has also been adopted in previous CLIP-based continual adaptation methods (Zhou et al., 2025b; Jha et al., 2024). We therefore preserve the textual space, freeze the original CLIP backbone, and train only a lightweight visual LoRA for each incoming task.

To learn composition-compatible representations, each task-specific LoRA is optimized using three complementary objectives. A text-alignment objective performs cross-entropy classification over

the current task's fixed textual anchors. A supervised contrastive objective alleviates visual feature crowding by pulling same-class samples together and pushing different classes apart. A negative-similarity objective suppresses responses between current-task images and previously observed textual anchors, thereby reducing cross-task interference. We additionally employ a prototype-initialized cumulative cosine classifier as an auxiliary visual branch for fine-grained recognition.

Explicitly composing all task-specific LoRA representations avoids task routing but requires one forward pass per task. To retain the compositional knowledge while improving inference efficiency, DISCIL introduces an incremental two-teacher LoRA consolidation strategy. The previous merged LoRA summarizes historical tasks, while the current LoRA captures newly acquired knowledge. Their features are combined to construct a teacher target and distilled into a new merged LoRA. This recursively consolidates historical and current knowledge without revisiting all previous modules.

Overall, DISCIL consists of three stages: task-specific visual adaptation, incremental LoRA consolidation, and task-agnostic inference. During inference, a test image passes through the frozen visual encoder and the final merged LoRA only once, after which its feature is classified using all seen-class textual anchors and the auxiliary cosine classifier. Consequently, DISCIL requires neither task identity, prompt retrieval, nor expert routing, while maintaining a single visual forward pass regardless of the number of tasks.

Our main contributions are summarized as follows:

- We exploit the distributive property of CLIP's inner-product classifier to reformulate task-specific PEFT from a module-routing problem into additive representation composition within a shared textual-anchor space.
- Motivated by the observed asymmetry between separated textual anchors and crowded visual features, we develop a visual-only PEFT strategy that learns selective and composition-compatible representations through text alignment, supervised contrastive learning, and negative similarity suppression.
- We propose an incremental two-teacher LoRA consolidation strategy that recursively compresses historical and current knowledge into a single merged LoRA, enabling rehearsal-free, router-free, and task-agnostic inference with only one visual forward pass.
- Extensive experiments on multiple benchmarks demonstrate that DISCIL achieves a favorable balance between continual recognition performance and inference efficiency.

## 2 Related Work

**Class-incremental learning**. CIL methods are commonly categorized into regularization-based (Li & Hoiem, 2017; Kirkpatrick et al., 2017), architecture-based (Mallya & Lazebnik, 2018; Fernando et al., 2017; Mallya et al., 2018), and replay-based (Rebuffi et al., 2017; Lopez-Paz & Ranzato, 2017) approaches. Although effective, dynamic expansion and replay generally introduce additional parameter or storage overhead. PEFT methods introduce a small number of task-specific parameters, such as prompts or adapters, while freezing most of the backbone (Wang et al., 2022b;a; Smith et al., 2023). Other methods, including MG-CLIP (Huang et al., 2025) and SLCA (Zhang et al., 2023), demonstrate that maintaining the stability of the pre-trained representation is important for downstream class-incremental learning. Our method follows this direction and further investigates how routing and retrieval can be replaced by direct composition of task-specific PEFT representations.

**Class-Incremental Learning with CLIP**. The transferable image-text representation of CLIP has recently been widely explored for CIL. Some methods introduce additional modules (Zhou et al., 2025b;a; Ma et al., 2025; Wang et al., 2023a) after the frozen CLIP encoders. Other approaches (Zheng et al., 2023; Yu et al., 2024; Marczak et al., 2024; Zhang et al., 2024) directly adapt CLIP parameters. Most existing approaches rely on feature replay, auxiliary data, expert selection, or task-wise model execution. In contrast, our method starts from the inner-product classification mechanism inherent to CLIP. By exploiting its distributive property, we formulate task-specific visual representations as additively composable contributions to shared text-anchor similarities. We further consolidate these representations into a single LoRA, enabling rehearsal-free and task-agnostic inference without prompt retrieval, task routing, or multiple visual forward passes.

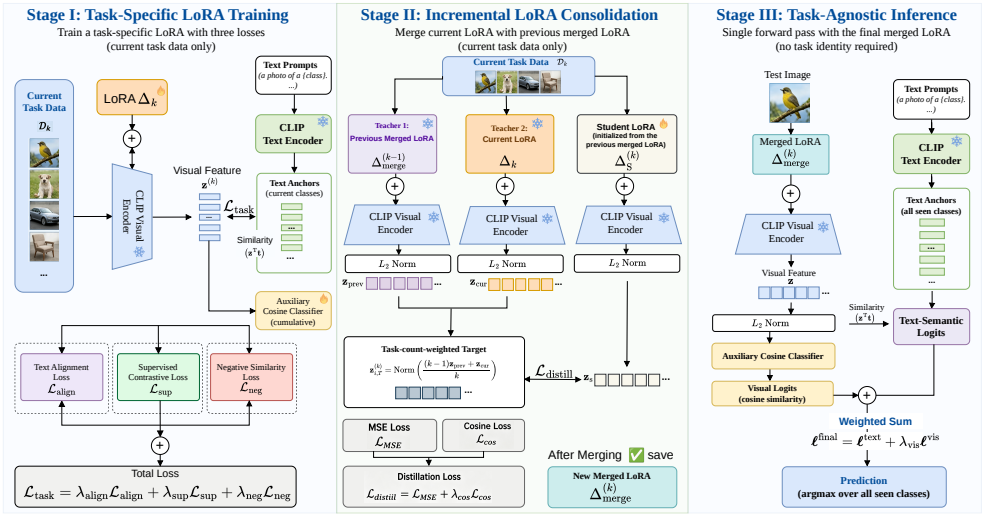


Figure 2: **Overall framework of DISCIL.** In Stage I, a task-specific visual LoRA is trained using text alignment, supervised contrastive learning, and negative similarity suppression. In Stage II, the current LoRA and the previous merged LoRA are consolidated through two-teacher feature distillation. In Stage III, the final merged LoRA enables task-agnostic classification with one visual forward pass using fixed text anchors and an auxiliary cumulative cosine classifier.

# 3 METHOD

## 3.1 PROBLEM FORMULATION AND OVERVIEW

We consider rehearsal-free CIL over a sequence of $K$ tasks. The training set of task $k$ is

$$\mathcal{D}_k = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_k}, \tag{5}$$

where $\mathbf{x}_i$ is an image and $y_i \in \mathcal{C}_k$ is its label. The class sets of different tasks are mutually disjoint.

When learning task $k$, only $\mathcal{D}_k$ is available. During inference, the task identity is unknown, and the model must classify each image among all classes observed so far:

$$\mathcal{C}_{1:k} = \bigcup_{j=1}^{k} \mathcal{C}_j. \tag{6}$$

Let $f_{\mathrm{v}}$ and $f_{\mathrm{t}}$ denote the CLIP visual and text encoders. For each class $c$, we construct $M$ textual prompts and average their embeddings to obtain stable text anchors:

$$\mathbf{t}_c = \mathrm{Norm}\left(\frac{1}{M}\sum_{m=1}^{M} f_{\mathrm{t}}\left(q_m(c)\right)\right), \tag{7}$$

where $q_m(c)$ is the $m$-th prompt template and $\mathrm{Norm}(\mathbf{a}) = \mathbf{a}/\|\mathbf{a}\|_2$.

As derived in Eqs. 2 and 3 before, the similarity between a composed visual representation and a text anchor can be decomposed into the additive contributions of individual task-wise representations. We use uniform composition weights, i.e., $\alpha_k = 1$ for all tasks.

As illustrated in Fig. 2, DISCIL consists of three stages: task-specific LoRA training, incremental LoRA consolidation, and task-agnostic inference.

## 3.2 STAGE I: TASK-SPECIFIC LORA TRAINING

For task $k$, we insert a task-specific LoRA $\Delta_k$ into the frozen CLIP visual encoder. Both CLIP encoders remain fixed, and only $\Delta_k$ is optimized. Given an image $\mathbf{x}_i$, we obtain

$$\mathbf{h}_i^{(k)} = f_{\mathrm{v}}(\mathbf{x}_i; \Delta_k), \qquad \mathbf{z}_i^{(k)} = \mathrm{Norm}\left(\mathbf{h}_i^{(k)}\right). \tag{8}$$

We use three complementary objectives to produce high responses to relevant text anchors, improve visual class separation, and suppress responses to unrelated old classes.

**Text Alignment Loss**. The similarities between the adapted visual feature and the current-task text anchors are used as classification logits:

$$\ell_{i,c}^{\text{text}} = (\mathbf{z}_i^{(k)\top}\mathbf{t}_c)/\tau_{\text{text}}, \qquad c \in \mathcal{C}_k, \tag{9}$$

where $\tau_{\text{text}}$ is a temperature parameter. The text alignment loss is

$$\mathcal{L}_{\text{align}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\mathbf{z}_i^{(k)\top}\mathbf{t}_{y_i}/\tau_{\text{text}}\right)}{\sum_{c\in\mathcal{C}_k}\exp\left(\mathbf{z}_i^{(k)\top}\mathbf{t}_c/\tau_{\text{text}}\right)}. \tag{10}$$

This objective increases the response to the GT text anchor and reduces the responses to competing current-class anchors, thereby performing classification while preserving image–text alignment.

**Supervised Contrastive Loss**. Because fine-grained visual features are often crowded, text alignment alone may not provide sufficiently clear visual boundaries. For an anchor sample $i$, we define its positive set as

$$\mathcal{P}(i) = \{p \neq i \mid y_p = y_i\}, \tag{11}$$

and the valid anchor set as

$$\mathcal{V} = \{i \mid |\mathcal{P}(i)| > 0\}. \tag{12}$$

Following the implementation, all positive similarities are aggregated before taking the logarithm:

$$\mathcal{L}_{\text{sup}} = -\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}\log\frac{\sum_{p\in\mathcal{P}(i)}\exp\left(\mathbf{z}_i^{(k)\top}\mathbf{z}_p^{(k)}/\tau_{\text{sup}}\right)}{\sum_{\substack{a=1\\a\neq i}}^{B}\exp\left(\mathbf{z}_i^{(k)\top}\mathbf{z}_a^{(k)}/\tau_{\text{sup}}\right)}, \tag{13}$$

where $\tau_{\text{sup}}$ is the contrastive temperature. Samples without a positive partner are excluded; when $\mathcal{V} = \varnothing$, we set $\mathcal{L}_{\text{sup}} = 0$. This loss reduces intra-class variation and improves inter-class separation.

**Negative Similarity Loss**. A task-specific representation may interfere with old classes if it produces high responses to their text anchors. Let $\mathcal{C}_{\text{old}}^k$ denote the set of all old classes before task $k$. We suppress the responses to all old text anchors using

$$\mathcal{L}_{\text{neg}} = \frac{1}{B|\mathcal{C}_{\text{old}}^k|}\sum_{i=1}^{B}\sum_{c\in\mathcal{C}_{\text{old}}^k}\left[\max\left(0, \mathbf{z}_i^{(k)\top}\mathbf{t}_c - m\right)\right]^2, \tag{14}$$

where $m$ is the similarity margin. We use $m = 0$, such that positive similarities to old text anchors are penalized. The loss considers all old anchors and is omitted for the first task.

The complete objective for current task-specific LoRA is

$$\mathcal{L}_{\text{task}} = \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{sup}}\mathcal{L}_{\text{sup}} + \lambda_{\text{neg}}\mathcal{L}_{\text{neg}}. \tag{15}$$

The three terms respectively encourage semantic alignment, visual discriminability, and low cross-task interference.

**Auxiliary Cumulative Cosine Classifier**. Although text anchors provide stable semantic references, they may not fully capture subtle visual differences between fine-grained classes. Inspired by Huang et al. (2025), we maintain a cumulative cosine classifier as an auxiliary branch.

After training $\Delta_k$, we use it to extract features for the current task. For each class $c \in \mathcal{C}_k$, its visual prototype is

$$\mathbf{p}_c = \text{Norm}\left(\frac{1}{N_c}\sum_{i:y_i=c}\mathbf{h}_i^{(k)}\right), \tag{16}$$

where features are averaged before normalization. The classifier is expanded for the new classes, whose weights are initialized as

$$\mathbf{w}_c \leftarrow \mathbf{p}_c, \qquad c \in \mathcal{C}_k. \tag{17}$$

The cosine-classifier logit for class $c$ is

$$\ell_{i,c}^{\text{vis}} = \gamma\frac{\mathbf{h}_i^{(k)\top}\mathbf{w}_c}{\|\mathbf{h}_i^{(k)}\|_2\|\mathbf{w}_c\|_2}, \tag{18}$$

where $\gamma$ is a learnable scale. The classifier is trained using task-local cross-entropy:

$$\mathcal{L}_{\text{cls}} = \text{CE}\left(\left\{\ell_{i,c}^{\text{vis}}\right\}_{c\in\mathcal{C}_k}, y_i\right). \tag{19}$$

Only current-task logits participate in this loss. During this stage, CLIP and the task-specific LoRA are fixed, and no historical data are used.

### 3.3 STAGE II: INCREMENTAL LORA CONSOLIDATION

Explicitly applying all task-specific LoRAs realizes the composition in Eq. 3, but requires one visual forward pass per LoRA. We therefore incrementally consolidate their knowledge into a single merged LoRA.

After the first task, the merged LoRA is initialized as $\Delta_{\text{merge}}^{(1)} = \Delta_1$.

For $k > 1$, the previous merged LoRA $\Delta_{\text{merge}}^{(k-1)}$ represents the first $k-1$ tasks, while $\Delta_k$ represents the current task. Given a current-task image $\mathbf{x}_i$, the two teacher branches produce

$$\mathbf{z}_{i,\text{prev}} = \text{Norm}\left(f_{\text{v}}\left(\mathbf{x}_i; \Delta_{\text{merge}}^{(k-1)}\right)\right), \mathbf{z}_{i,\text{cur}} = \text{Norm}\left(f_{\text{v}}\left(\mathbf{x}_i; \Delta_k\right)\right). \tag{20}$$

Because the previous merged LoRA summarizes $k-1$ tasks and the current LoRA represents one task, we construct the task-count-weighted teacher target as

$$\mathbf{z}_{i,T}^{(k)} = \text{Norm}\left(\frac{(k-1)\mathbf{z}_{i,\text{prev}} + \mathbf{z}_{i,\text{cur}}}{k}\right). \tag{21}$$

The effective relative contribution of the two branches is $(k-1) : 1$.

A student LoRA is initialized from the previous merged LoRA: $\Delta_{\text{S}}^{(k)} \leftarrow \Delta_{\text{merge}}^{(k-1)}$.

Its normalized output is

$$\mathbf{z}_{i,S}^{(k)} = \text{Norm}\left(f_{\text{v}}\left(\mathbf{x}_i; \Delta_{\text{S}}^{(k)}\right)\right). \tag{22}$$

Let $D$ denote the feature dimension. The student is optimized using following distillation objective:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{BD}\sum_{i=1}^{B}\sum_{d=1}^{D}\left(z_{i,S,d}^{(k)} - z_{i,T,d}^{(k)}\right)^2, \mathcal{L}_{\cos} = \frac{1}{B}\sum_{i=1}^{B}\left[1 - \cos\left(\mathbf{z}_{i,S}^{(k)}, \mathbf{z}_{i,T}^{(k)}\right)\right], \tag{23}$$

and the whole distillation loss is

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{MSE}} + \lambda_{\cos}\mathcal{L}_{\cos}. \tag{24}$$

We use $\lambda_{\cos} = 0.5$. Both terms match the student to the same normalized teacher target and jointly define the feature-distillation objective.

After convergence, the student becomes the new merged LoRA: $\Delta_{\text{merge}}^{(k)} = \Delta_{\text{S}}^{(k)}$.

This stage uses only current-task images and does not include classification or the Stage-I losses. The separate current-task LoRA can then be discarded. Each consolidation step therefore requires only the previous merged LoRA and the current LoRA, rather than all historical LoRAs.

Note that although the recursive consolidation is not an exact algebraic realization of Eq. 3, it is designed to approximate the same explicit composition target and preserve its resulting classification behavior. As verified in subsequent experiments, the consolidated single-LoRA closely matches its incremental composition target and achieves downstream performance comparable to explicit multi-LoRA composition, while significantly improving inference efficiency.

### 3.4 STAGE III: TASK-AGNOSTIC INFERENCE

After learning task $k$, only the merged LoRA $\Delta_{\text{merge}}^{(k)}$ is required. For a test image $\mathbf{x}$, we perform one visual forward pass:

$$\mathbf{h} = f_{\text{v}}\left(\mathbf{x}; \Delta_{\text{merge}}^{(k)}\right), \qquad \mathbf{z} = \text{Norm}(\mathbf{h}). \tag{25}$$

The text-semantic branch compares $\mathbf{z}$ with all seen-class text anchors:

$$\boldsymbol{\ell}^{\text{text}} = \gamma_{\text{text}}\mathbf{z}\mathbf{T}_{1:k}^{\top}, where\ \mathbf{T}_{1:k} = \left[\mathbf{t}_1, \ldots, \mathbf{t}_{|\mathcal{C}_{1:k}|}\right]. \tag{26}$$

The auxiliary visual branch produces

$$\boldsymbol{\ell}^{\text{vis}} = \text{CosineClassifier}(\mathbf{h}). \tag{27}$$

Table 1: Performance comparison among different methods. Each dataset is split into 10 tasks.

| Methods | CIFAR100 | | ImageNet-R | | CUB | | Aircraft | | Cars | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Avg | Last | Avg | Last | Avg | Last | Avg | Last | Avg | Last |
| ZS-CLIP | 81.38 | 71.26 | 82.93 | 76.67 | 76.01 | 64.50 | 26.66 | 17.22 | 82.60 | 76.37 |
| CoOp | 83.37 | 73.36 | 82.40 | 76.20 | 77.34 | 68.70 | 44.26 | 39.87 | 89.73 | 84.91 |
| L2P | 76.42 | 66.21 | 75.73 | 67.22 | 79.23 | 68.54 | 55.06 | 44.88 | 83.81 | 72.44 |
| DualPrompt | 79.07 | 70.06 | 78.47 | 70.82 | 83.21 | 74.94 | 50.93 | 46.53 | 85.30 | 74.35 |
| CODA-P | 83.53 | 72.20 | 74.45 | 65.88 | 83.36 | 76.12 | 54.51 | 45.06 | 96.16 | 92.17 |
| PROOF | 86.70 | 79.05 | 85.34 | 80.10 | 84.93 | 79.43 | 61.00 | 53.59 | 93.26 | 89.84 |
| CLAP | 74.19 | 63.45 | 81.22 | 75.80 | 81.55 | 73.86 | 55.32 | 45.17 | 92.18 | 90.34 |
| TPPT-V | 87.82 | 79.31 | 84.93 | 80.83 | 85.88 | 80.34 | 63.56 | 55.15 | 95.12 | 94.13 |
| MG-CLIP | 87.12 | 80.13 | 87.13 | 81.93 | 85.12 | 79.86 | 59.83 | 49.89 | 93.21 | 90.06 |
| DISCIL (Ours) | 88.34 | 80.86 | 87.75 | 82.54 | 85.97 | 80.23 | 62.13 | 54.13 | 95.23 | 93.43 |

The two branches contain the same number of logits and are fused as

$$\boldsymbol{\ell}^{\text{final}} = \boldsymbol{\ell}^{\text{text}} + \lambda_{\text{vis}} \boldsymbol{\ell}^{\text{vis}}. \tag{28}$$

$\lambda_{\text{vis}}$ controls ratio between the two. The final prediction is

$$\hat{y} = \arg\max_{c \in \mathcal{C}_{1:k}} \ell_c^{\text{final}}. \tag{29}$$

The inference process requires neither task identity, prompt retrieval, query–key matching, nor expert routing. Since all task knowledge is recursively consolidated into one LoRA, the number of visual forward passes remains one regardless of the number of learned tasks.

# 4 EXPERIMENTS

## 4.1 EXPERIMENT SETTINGS

**Datasets.** Following (Lu et al., 2025), we evaluate our method through CIL tasks on five benchmark datasets: CIFAR-100 (Krizhevsky et al., 2009), ImageNet-R (Hendrycks et al., 2021), CUB-200 (Wah et al., 2011), FGVC-Aircraft (Maji et al., 2013) and Stanford Cars (Krause et al., 2013).

**Evaluation Metrics.** Following most previous works (Lu et al., 2025; Huang et al., 2025), let $A_t$ denote the average test accuracy across the first $t$ tasks after the model has completed training on the $t$-th task. We evaluate performance using two key metrics: **Avg**: The overall mean accuracy across all historical tasks up to the current step, defined as $\frac{1}{K}\sum_{i=1}^{K} A_i$. **Last**: The final average accuracy obtained after completing the training of the last task $K$, represented as $A_K$.

**Implementation Details.** We implement our method in PyTorch on a single NVIDIA RTX 5090 GPU. We use a batch size of 64 and insert task-specific LoRA modules ($r = 4$, $\alpha = 8$, and zero dropout) into the attention output projection and two MLP projections of each visual Transformer block. More detailed information can be found in Appendix C.

## 4.2 COMPARISON RESULTS

As shown in Tab. 1, we compare the zero-shot CLIP baseline (Radford et al., 2021), the prompt-tuning baseline CoOp (Zhou et al., 2022), representative prompt-based continual learning methods including L2P (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), and CODA-P (Smith et al., 2023), as well as recent CLIP-based CIL approaches such as PROOF (Zhou et al., 2025b), CLAP (Jha et al., 2024), TPPT-V (Lu et al., 2025), and MG-CLIP (Huang et al., 2025). These methods cover different continual adaptation strategies, including zero-shot transfer, prompt selection, cross-modal adaptation, and feature-space refinement.

Overall, DISCIL achieves strong and consistent performance across all five benchmarks, including both general recognition datasets and challenging fine-grained datasets. Compared with prompt-based methods that rely on task-specific prompt retrieval or selection, DISCIL benefits from directly composing task-wise visual representations in the shared CLIP text-anchor space. It also remains competitive with recent CLIP-based approaches that introduce additional cross-modal modules or specialized adaptation mechanisms. The favorable average accuracy indicates that DISCIL maintains stable performance throughout the incremental process, while the strong final accuracy shows

Table 2: **Generalization of DISCIL across different PEFT modules.** We instantiate DISCIL with visual prompts, adapters, and LoRA (default setting in this paper).

| PEFT Module | CIFAR-100 | | ImageNet-R | |
|---|---|---|---|---|
| | Avg. | Last | Avg. | Last |
| Prompt | 87.23 | 80.12 | 87.02 | 81.93 |
| Adapter | 87.89 | 80.82 | 87.53 | 81.73 |
| LoRA | 88.34 | 80.86 | 87.75 | 82.54 |

Table 3: **Comparison of LoRA composition and consolidation strategies.** We report the number of visual-encoder forward passes per image during LoRA consolidation and task-agnostic inference, together with continual learning performance. Here, $k$ denotes the number of tasks learned so far.

| Strategy | Forward Passes | | CIFAR-100 | | ImageNet-R | |
|---|---|---|---|---|---|---|
| | Consolidation | Inference | Avg. | Last | Avg. | Last |
| EMC | – | $k$ | 88.52 | 80.57 | 87.51 | 82.34 |
| ALD | $k+1$ | 1 | 88.42 | 80.41 | 87.27 | 82.73 |
| **DISCIL** | **3** | **1** | 88.34 | 80.86 | 87.75 | 82.54 |

that it effectively preserves previously learned knowledge after all tasks are completed. These results validate the effectiveness and generality of the proposed distributive similarity composition strategy. Moreover, incrementally consolidating task-specific LoRAs into a single merged LoRA retains competitive continual learning performance while avoiding replay, task routing, and multiple visual forward passes during inference.

### 4.3 TASK-SELECTIVE LORA RESPONSES

As shown in Fig. 3, the diagonal entry in each row is consistently the largest, indicating that images from task $i$ obtain the highest GT-anchor similarity when processed by the matched LoRA. In contrast, unmatched LoRAs produce relatively lower responses. This task-selective response pattern allows the matched task representation to contribute more strongly to the correct classes while limiting unrelated contributions during feature composition, thereby providing empirical support for our distributive similarity composition principle.

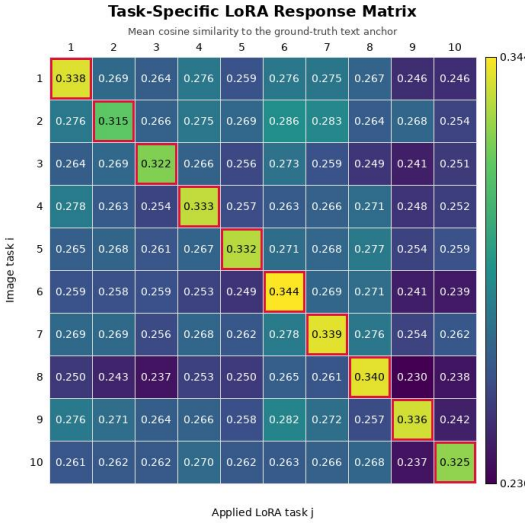


Figure 3: **Task-wise response matrix.** Each entry denotes the average cosine similarity between the GT text anchors of images from task $i$ and the visual features produced by task-specific LoRA $j$. The consistently higher diagonal responses show that the matched LoRA produces stronger responses to its corresponding task than unmatched LoRAs.

### 4.4 GENERALIZATION TO DIFFERENT PEFT MODULES

Although DISCIL is instantiated with LoRA in main experiments, its core principle is not restricted to a specific PEFT architecture. To verify this, we replace the LoRA with task-specific Prompts and Adapters, while keeping the frozen CLIP backbone, fixed text anchors, training objectives, task splits, and evaluation protocol unchanged. The same feature-level consolidation strategy is used to compress the task-specific representations into a single module for inference. Specific implementation details for Prompt and Adapter can be found in Appendix C.

As shown in Tab. 2, the prompt- and adapter-based variants also achieve competitive continual learning performance on both datasets. This indicates that the effectiveness of DISCIL does not rely on the specific parameterization of LoRA. Instead, the key lies in learning task-selective visual representations and composing their contributions within the shared CLIP text-anchor space. These results demonstrate that distributive similarity composition can be generalized to different visual PEFT modules, while LoRA is adopted as the default implementation in our main experiments.

### 4.5 ABLATION STUDY

**Comparison of LoRA Composition and Consolidation Strategies.** Tab. 3 compares three strategies for combining task-specific LoRA representations. Explicit Multi-LoRA Composition (EMC)

directly aggregates the features produced by all LoRAs and therefore requires $k$ visual forward passes during inference after learning $k$ tasks. All-LoRA Distillation (ALD) compresses the outputs of all LoRAs into a single LoRA, reducing inference to one forward pass, but its consolidation cost increases with the number of tasks. In contrast, DISCIL uses only the previous merged LoRA and the current task-specific LoRA as two teachers. Together with the student branch, each consolidation step requires only three forward passes, while inference always requires a single forward pass.

The three strategies achieve comparable continual-learning performance on both CIFAR-100 and ImageNet-R. EMC provides the direct implementation of the explicit feature composition in Eq. 3, whereas DISCIL approximates this multi-module composition using a single merged LoRA. Although the recursive consolidation is not an exact algebraic realization of Eq. 3, our optimization diagnostics show that the consolidated student closely matches its composition target, with student–teacher cosine similarities above $0.997$ on CIFAR-100 and nearly $1.0$ on ImageNet-R. Moreover, DISCIL achieves performance comparable to, and in several metrics slightly better than EMC, while reducing inference from $k$ visual forward passes to only one. These results suggest that the proposed consolidation provides an effective and efficient approximation to explicit representation composition. A more detailed analysis is provided in the Appendix D.

**Ablation of Three Losses in Stage I.** Tab. 4 evaluates both the effectiveness of three objectives and the sensitivity to their loss coefficients. The text alignment loss is the primary supervision for task-specific LoRA training: it directly performs classification over current textual anchors while preserving CLIP's image–text alignment. We therefore fix $\lambda_{\text{align}} = 1.0$ as the reference coefficient and tune the two auxiliary objectives relative to it.

Table 4: **Effectiveness and weight sensitivity of the three training objectives.** We fix the coefficient of the primary text alignment loss to $\lambda_{\text{align}} = 1.0$ and vary the weights of the supervised contrastive loss and negative similarity loss.

| Loss Weights | | | CIFAR-100 | | ImageNet-R | |
|---|---|---|---|---|---|---|
| $\lambda_{\text{align}}$ | $\lambda_{\text{sup}}$ | $\lambda_{\text{neg}}$ | Avg. | Last | Avg. | Last |
| 1.0 | 0.0 | 0.0 | 86.92 | 78.54 | 86.41 | 80.67 |
| 1.0 | 0.3 | 0.0 | 87.89 | 80.02 | 87.21 | 81.76 |
| 1.0 | 0.0 | 0.2 | 87.51 | 79.63 | 86.98 | 81.42 |
| 1.0 | 0.1 | 0.2 | 87.96 | 80.21 | 87.33 | 81.93 |
| 1.0 | 0.2 | 0.2 | 88.17 | 80.57 | 87.58 | 82.28 |
| **1.0** | **0.3** | **0.2** | **88.34** | **80.86** | **87.75** | **82.54** |
| 1.0 | 0.4 | 0.2 | 88.11 | 80.48 | 87.61 | 82.31 |
| 1.0 | 0.3 | 0.1 | 88.09 | 80.55 | 87.57 | 82.29 |
| 1.0 | 0.3 | 0.3 | 88.14 | 80.44 | 87.52 | 82.20 |

Using only the text alignment loss already provides a strong baseline, confirming its central role in overall learning. Adding the supervised contrastive loss consistently improves performance by reducing the crowding of fine-grained visual features and enlarging inter-class margins. The negative similarity loss also brings clear improvements by suppressing undesired responses to previously observed textual anchors, thereby reducing cross-task interference. Combining all three objectives achieves the strongest overall performance, showing that semantic alignment, visual discrimination, and interference suppression are complementary.

We further vary $\lambda_{\text{sup}}$ and $\lambda_{\text{neg}}$ around the default setting. The results remain stable within a reasonable range, while overly weak auxiliary losses provide insufficient regularization and overly strong weights may interfere with the primary text-alignment objective. Based on this study, we set $\lambda_{\text{sup}} = 0.3$ and $\lambda_{\text{neg}} = 0.2$ in all other experiments.

## 5 CONCLUSION

We presented DISCIL, a rehearsal-free class-incremental learning framework that replaces task-specific module routing with additive representation composition. By exploiting the distributive property of CLIP's inner-product classifier, DISCIL interprets task-wise visual representations as complementary contributions to shared text-anchor similarities. To support effective composition, task-specific visual LoRAs are trained with text alignment, supervised contrastive learning, and negative similarity suppression. An incremental two-teacher consolidation strategy further compresses historical and current knowledge into a single merged LoRA, enabling task-agnostic inference with only one visual forward pass. Extensive experiments demonstrate that DISCIL achieves strong continual recognition performance while maintaining parameter and inference efficiency, without replay, task identification, or expert routing.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with paper writing and language polishing, literature search and related-work organization, and software implementation. In particular, generative AI was used to improve the clarity, grammar, and readability of the manuscript, to help identify potentially relevant literature, and to assist in drafting and debugging parts of the experimental code. All AI-assisted outputs were carefully reviewed and verified by the authors. Relevant literature suggested by AI tools was manually checked against the original publications, and AI-assisted code was inspected, executed, and tested for correctness before being used in our experiments. The authors take full responsibility for the final content of this work, including all text, claims, experimental results, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work focuses on methodological research in class-incremental learning with pre-trained vision-language models. Our experiments are conducted on standard publicly available image-classification benchmarks and do not involve human subjects, collection of private or sensitive information, or the release of a new dataset. The proposed method inherits the limitations and potential biases of the underlying pre-trained CLIP model and the datasets on which it is evaluated. Therefore, its predictions should not be assumed to be free from demographic, cultural, or representation biases, and additional evaluation would be necessary before deployment in safety-critical or high-stakes applications. The main goal of this work is to improve the effectiveness and efficiency of continual adaptation, and we do not identify any additional ethical risks beyond those generally associated with the use of large-scale pre-trained vision-language models.

### REPRODUCIBILITY STATEMENT

We have made efforts to provide sufficient information for reproducing the proposed method and experiments. The main paper describes the complete DISCIL framework, including the task-specific LoRA training objectives, the incremental two-teacher consolidation procedure, the inference formulation, the datasets, evaluation metrics, and the main optimization settings. The supplementary material provides additional implementation details, including the construction of text anchors, PEFT configurations, loss temperatures and inference scales, as well as further analyses of the consolidation behavior and auxiliary classifier. Together, these materials specify the major model, training, and evaluation choices used to obtain the reported results.

## REFERENCES

Haoran Chen, Ping Wang, Zihan Zhou, Xu Zhang, Zuxuan Wu, and Yu-Gang Jiang. Achieving more with less: Additive prompt tuning for rehearsal-free class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 340–349, 2025.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

Saurabh Garg, Mehrdad Farajtabar, Hadi Pouransari, Raviteja Vemulapalli, Sachin Mehta, Oncel Tuzel, Vaishaal Shankar, and Fartash Faghri. Tic-clip: Continual training of clip models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=TLADT8Wrhn.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 2208–2222, 2021.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A criti-

cal analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Linlan Huang, Xusheng Cao, Haori Lu, and Xialei Liu. Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In *European Conference on Computer Vision*, pp. 214–231. Springer, 2024a.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3777–3786, 2025.

Wei-Cheng Huang, Chun-Fu Chen, and Hsiang Hsu. Ovor: Oneprompt with virtual outlier regularization for rehearsal-free class-incremental learning. *arXiv preprint arXiv:2402.04129*, 2024b.

Saurav Jha, Dong Gong, and Lina Yao. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *Advances in neural information processing systems*, 37:129146–129186, 2024.

Youngeun Kim, Yuhang Li, and Priyadarshini Panda. One-stage prompt-based continual learning. In *European conference on computer vision*, pp. 163–179. Springer, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Muhammad Rifki Kurniawan, Xiang Song, Zhiheng Ma, Yuhang He, Yihong Gong, Yang Qi, and Xing Wei. Evolving parameterized prompt memory for continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13301–13309, 2024.

Lan Li, Tao Hu, Da-Wei Zhou, Jia-Qi Yang, Han-Jia Ye, and De-Chuan Zhan. Bofa: Bridge-layer orthogonal low-rank fusion for clip-based class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22967–22975, 2026.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-CLIP: Multimodal continual learning for vision-language model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sb7qHFYwBc.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Haodong Lu, Xinyu Zhang, Kristen Moore, Jason Xue, Lina Yao, Anton van den Hengel, and Dong Gong. Continual learning on clip via incremental prompt tuning with intrinsic textual anchors. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=YJnjkzKq5Y.

Mao-Lin Luo, Zi-Hao Zhou, Tong Wei, and Min-Ling Zhang. LADA: Scalable label-specific CLIP adapter for continual learning. In *Forty-second International Conference on Machine Learning*, 2025.

Yue Ma, Huantao Ren, Boyu Wang, Jingang Jin, Senem Velipasalar, and Qinru Qiu. Lvp-clip: Revisiting clip for continual learning with label vector pool. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 231–240. IEEE, 2025.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 67–82, 2018.

Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. Magmax: Leveraging model merging for seamless continual learning. In *European Conference on Computer Vision*, pp. 379–395. Springer, 2024.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11909–11919, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Runqi Wang, Xiaoyue Duan, Guoliang Kang, Jianzhuang Liu, Shaohui Lin, Songcen Xu, Jinhu Lü, and Baochang Zhang. Attriclip: A non-incremental learner for incremental knowledge learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3654–3663, 2023a.

Yabin Wang, Zhiheng Ma, Zhiwu Huang, Yaowei Wang, Zhou Su, and Xiaopeng Hong. Isolation and impartial aggregation: A paradigm of incremental learning without interference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10209–10217, 2023b.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pp. 631–648. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022b.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 374–382, 2019.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19091–19101, 2023. doi: 10.1109/ICCV51070.2023.01754.

Wentao Zhang, Yujun Huang, Weizhuo Zhang, Tong Zhang, Qicheng Lao, Yue Yu, Wei-Shi Zheng, and Ruixuan Wang. Continual learning of image classes with language guidance from a vision-language model. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12): 13152–13163, 2024.

Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19125–19136, 2023.

Da-Wei Zhou, Kai-Wen Li, Jingyi Ning, Han-Jia Ye, Lijun Zhang, and De-Chuan Zhan. External knowledge injection for clip-based class-incremental learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3314–3325. IEEE, 2025a.

Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4489–4504, 2025b.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.

## A Cross-Dataset Visual–Text Geometry and Design Implications

To further examine the feature-space observation presented in Fig. 1(b) of the main paper, we visualize CLIP image and text representations across all five downstream benchmarks and a public web image–text dataset. For each downstream dataset, we randomly sample 512 images and construct the corresponding text features using their class-name prompts. For Conceptual Captions, we sample 512 naturally paired images and free-form captions. All image and text features are extracted using a frozen CLIP model, $\ell_2$-normalized, and jointly projected into two dimensions using UMAP (McInnes et al., 2018) with cosine distance. UMAP is fitted independently for each dataset; therefore, the absolute coordinates and orientations should not be compared across different panels.

**Visual crowding on downstream classification datasets.** As shown in Fig. 4(a)–(e), the frozen visual features consistently occupy relatively compact regions, whereas the class-name text features form more separated semantic anchors. This observation is consistent with the mean pairwise cosine similarities reported in the main paper, where visual features exhibit higher intra-modal similarity than their corresponding text anchors. The UMAP visualizations provide qualitative evidence of this asymmetry, while the cosine statistics measure it directly in the original high-dimensional feature space.

The crowded visual geometry can be attributed to the large amount of instance-level variation contained in images. Images from either the same or different classes may vary substantially in background, viewpoint, scale, illumination, pose, and appearance. This issue is especially pronounced

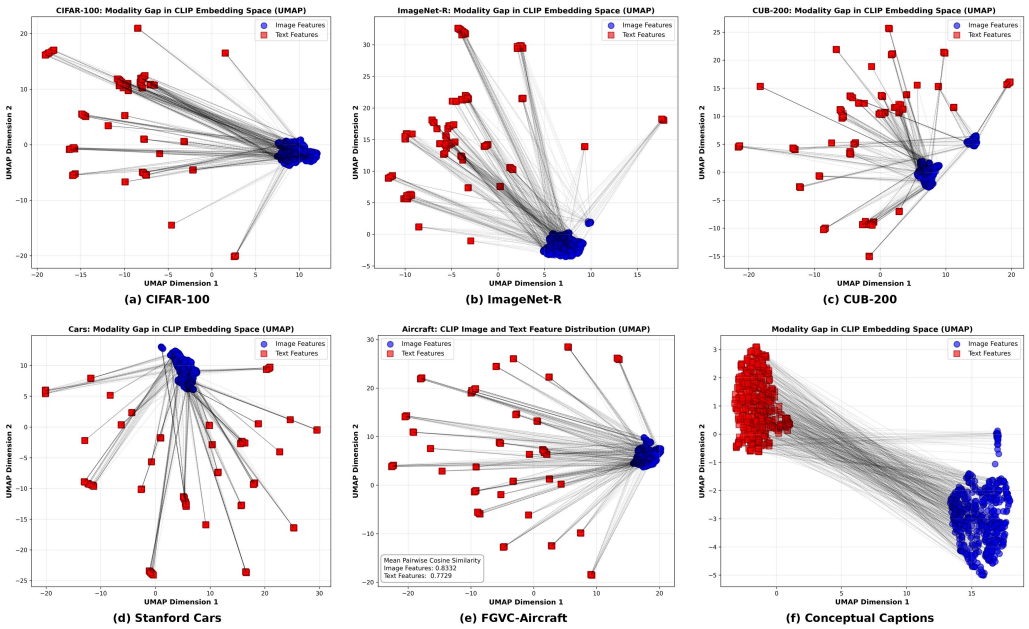


Figure 4: **Cross-dataset visualization of CLIP's visual–text feature geometry.** Panels (a)–(e) show the five downstream classification benchmarks, where each image is paired with a text feature constructed from its class-name prompt. Panel (f) shows naturally paired web images and free-form captions from Conceptual Captions (Sharma et al., 2018). Blue circles denote image features, red squares denote text features, and each gray line connects a matched image–text pair. On downstream classification datasets, visual features occupy relatively compact regions, whereas class-name text anchors are distributed across more separated semantic locations. In contrast, free-form captions form a denser text manifold because nearly every image is associated with a different sentence.

on fine-grained datasets such as CUB-200, Stanford Cars, and FGVC Aircraft, where different categories are often distinguished only by subtle local details. Although CLIP pre-training encourages semantic compatibility between matched images and texts, it does not explicitly enforce large classification margins between all possible downstream visual categories. Consequently, frozen CLIP features can preserve broad semantic information while compressing the fine-grained distinctions required by downstream recognition.

**Stable class-level textual references.** The text inputs used for downstream classification have substantially different characteristics. They are clean class names embedded through fixed prompt templates and contain little instance-level nuisance variation. Samples from the same category therefore share the same class-level text representation, allowing the text features to serve as stable semantic prototypes.

Moreover, the CLIP text encoder has been pre-trained on large-scale and semantically diverse image–text data, providing a structured semantic space for a broad range of class concepts. During continual learning, both the text encoder and the prompt templates remain fixed. The resulting text anchors are therefore unchanged across incremental stages and provide consistent references for both previously observed and newly introduced classes. Based on this observation, we freeze the text encoder and adapt only the visual encoder through task-specific PEFT modules.

**Downstream class prompts versus natural captions.** The geometry of Conceptual Captions in Fig. 4(f) differs from that of the downstream classification datasets. In panels (a)–(e), the text features correspond to a limited set of repeated class-name prompts. Consequently, many samples share exactly the same text representation, producing a sparse collection of class-level anchors. In contrast, almost every web image is paired with a different free-form sentence in Conceptual Captions.

These captions describe diverse objects, attributes, actions, and contexts, resulting in a substantially denser and more varied text manifold.

Panel (f) also shows a visible separation between the image and text distributions. This is consistent with the intrinsic modality gap of CLIP: the two encoders process different input structures and are not explicitly optimized to make their marginal feature distributions identical. Such separation does not imply unsuccessful image–text alignment, because CLIP classification depends on relative cosine similarities in the original high-dimensional space. In addition, UMAP is a nonlinear projection and does not preserve all global distances. Instead, this result indicates that modality-dependent geometry already exists in natural web image–text data and is not caused solely by downstream domain shift.

**Connection to the design of DISCIL.** The observed visual–text asymmetry directly motivates the visual-only adaptation strategy of DISCIL. We preserve the stable textual space and use its class anchors as fixed cross-task references, while adapting the more crowded visual representations. The three Stage-I objectives address different aspects of this geometry. The text-alignment loss $\mathcal{L}_{\text{align}}$ increases the similarity to the ground-truth text anchor while reducing similarities to competing current-task anchors, thereby maintaining CLIP's image–text alignment. The supervised contrastive loss $\mathcal{L}_{\text{sup}}$ directly alleviates visual crowding by reducing intra-class variation and enlarging inter-class margins. The negative-similarity loss $\mathcal{L}_{\text{neg}}$ suppresses undesired responses to previously observed text anchors, encouraging task-selective representations and reducing interference during additive composition.

Nevertheless, class-name text anchors primarily encode linguistic semantics and may not fully describe the subtle appearance differences required for fine-grained recognition. Following MG-CLIP (Huang et al., 2025), we therefore adopt a prototype-initialized cumulative cosine classifier as an auxiliary visual branch. Its classifier weights are constructed from training-image features and consequently provide decision boundaries directly grounded in the visual space. The two branches are complementary: the text-anchor branch supplies stable and transferable semantic references, whereas the auxiliary visual classifier captures appearance-level distinctions that may be difficult to express using class names alone.

## B Extended Related Work

The transferable image–text representations of CLIP (Radford et al., 2021) have recently been widely explored for class-incremental learning. Existing methods mainly differ in how they adapt the pre-trained representation, preserve previously learned knowledge, and integrate task-specific information during inference.

A common direction freezes the CLIP backbone and introduces lightweight modules for incremental adaptation. AttriCLIP (Wang et al., 2023a) constructs category prompts from a learnable attribute bank, while PROOF (Zhou et al., 2025b) expands task-specific projection modules and employs cross-modal fusion. LVP-CLIP (Ma et al., 2025) maintains image-derived label vectors instead of relying only on textual labels, and LADA (Luo et al., 2025) introduces scalable label-specific adapters with feature distillation. These methods preserve the pre-trained backbone but generally require additional task- or class-specific modules.

Another line of work performs prompt or adapter tuning within CLIP. CLAP4CLIP (Jha et al., 2024) adopts probabilistic fine-tuning and distribution regularization to transfer pre-trained knowledge across tasks. MOE4CL (Yu et al., 2024) dynamically expands mixture-of-experts adapters and selects suitable experts during inference. TPPT (Lu et al., 2025) uses intrinsic textual anchors to guide incremental prompt tuning, whereas RAPF (Huang et al., 2024a) adjusts class representations according to textual relationships and subsequently fuses task-specific parameters. Although effective, several of these approaches rely on prompt retrieval, expert selection, or parameter fusion to determine how task-specific knowledge should be used.

Recent studies further investigate the geometry and knowledge structure of CLIP. MG-CLIP (Huang et al., 2025) preserves and compensates for the vision–language modality gap during continual adaptation. BOFA (Li et al., 2026) performs orthogonal low-rank fusion in the cross-modal bridge layer and combines visual and textual prototypes. ENGINE (Zhou et al., 2025a) enhances both modalities with external visual and textual knowledge. ZSCL (Zheng et al., 2023) focuses on preserving CLIP's

zero-shot transfer ability, while MagMax (Marczak et al., 2024) merges task vectors in parameter space to accumulate sequential knowledge.

In contrast to these methods, DISCIL starts from CLIP's inner-product classification mechanism. We interpret the outputs of task-specific PEFT modules as complementary visual representations whose similarities can be additively composed in the shared text-anchor space. The task-wise representations are then incrementally consolidated into a single module, avoiding task identification, prompt retrieval, expert routing, and multi-module inference.

## C Additional Implementation Details

**Implementation Details** We implement our method in PyTorch on a single NVIDIA RTX 5090 GPU. We use a batch size of 64 and insert task-specific LoRA modules ($r = 4$, $\alpha = 8$, and zero dropout) into the attention output projection and two MLP projections of each visual Transformer block. Each task is trained for 5 epochs using SGD with an initial learning rate of $10^{-3}$, momentum $0.9$, and weight decay $5 \times 10^{-4}$. We employ a 10% linear warm-up followed by cosine decay to $10^{-5}$. The coefficients of text-alignment, supervised contrastive, and negative-similarity losses are $1$, $0.3$, and $0.2$, respectively. $\lambda_{\text{vis}}$ is set to $1$. Incremental LoRA consolidation is performed for 5 epochs using Adam with a learning rate of $10^{-3}$. The prototype-initialized cosine classifier is trained for 5 epochs with a learning rate of $5 \times 10^{-4}$ and logit scale $\gamma$ is initialized to $10$.

**Implementation Details of Different PEFT Instantiations** To examine whether the proposed composition principle depends on a particular PEFT architecture, we instantiate DISCIL with LoRA, additive visual prompts, and adapters. Across all variants, the CLIP backbone and text anchors remain frozen, while only the task-specific PEFT parameters are optimized using the three Stage-I objectives. After each task, the previous merged module and the current task-specific module serve as two teachers, and their composed feature target is distilled into a new merged module. Therefore, all three variants follow the same training, consolidation, and single-pass inference pipeline; only the PEFT parameterization is changed.

**LoRA.** For the default implementation, we insert LoRA modules with rank $r = 4$, scaling factor $\alpha = 8$, and zero dropout into the attention output projection and the two MLP projections of every visual Transformer block. The CLIP parameters remain frozen, and only the newly introduced LoRA parameters are updated. During consolidation, the student LoRA has the same architecture and is initialized from the previous merged LoRA.

**Prompt.** Following the additive prompt parameterization of APT (Chen et al., 2025), prompts are incorporated directly into the attention computation of the CLS token rather than concatenated with the input token sequence. Different from the original shared-prompt setting, we learn a task-specific set of additive visual prompts for each incoming task. Each task uses the same fixed number of prompt vectors, and only these prompt parameters are optimized during Stage I. During consolidation, the previous merged prompts and the current task-specific prompts produce the two teacher representations. A student prompt module, initialized from the previous merged prompts, is then optimized using the same feature-distillation objective and retained for subsequent tasks.

**Adapter.** For the adapter variant, a residual bottleneck adapter is inserted after both the self-attention sublayer and the MLP sublayer of every visual Transformer block. The bottleneck dimension is selected such that the overall number of trainable adapter parameters is approximately matched to that of the LoRA configuration. For each task, only the corresponding adapter parameters are trained, while the original CLIP backbone remains fixed. The previous merged adapter and the current task-specific adapter are subsequently consolidated into a student adapter using the same two-teacher feature-distillation procedure.

These implementations maintain comparable parameter budgets and identical task splits, textual anchors, training objectives, and evaluation protocols. The results in Table 2 of the main paper therefore demonstrate that the effectiveness of DISCIL arises from task-wise representation composition and consolidation, rather than from the specific structure of LoRA.

**Text-anchor construction.** Following the prompt-engineering and prompt-ensembling protocol introduced by CLIP (Radford et al., 2021), we adopt the dataset-specific templates released in

the official CLIP implementation. The same template protocol has also been adopted by subsequent CLIP adaptation and continual learning methods (Lu et al., 2025). We inherit these established templates rather than selecting them through validation on the incremental tasks. For every class $c$, we substitute its canonical class name for `[CLASS]` in each dataset-specific template, encode all resulting prompts using the frozen text encoder, average their unnormalized text features, and finally apply $\ell_2$ normalization. The templates and the corresponding number $M$ are as follows. For CIFAR-100, $M = 18$, and we use `a photo of a [CLASS].`, `a blurry photo of a [CLASS].`, `a black and white photo of a [CLASS].`, `a low contrast photo of a [CLASS].`, `a high contrast photo of a [CLASS].`, `a bad photo of a [CLASS].`, `a good photo of a [CLASS].`, `a photo of a small [CLASS].`, `a photo of a big [CLASS].`, and the corresponding nine templates obtained by replacing "a [CLASS]" with "the [CLASS]". For ImageNet-R, $M = 7$, and the templates are `itap of a [CLASS].`, `a bad photo of the [CLASS].`, `a origami [CLASS].`, `a photo of the large [CLASS].`, `a [CLASS] in a video game.`, `art of the [CLASS].`, and `a photo of the small [CLASS].`. For CUB-200, $M = 1$, using `a photo of a [CLASS].`. For FGVC-Aircraft, $M = 2$, using `a photo of a [CLASS], a type of aircraft.` and `a photo of the [CLASS], a type of aircraft.`. For Stanford Cars, $M = 8$, using `a photo of a [CLASS].`, `a photo of the [CLASS].`, `a photo of my [CLASS].`, `i love my [CLASS]!`, `a photo of my dirty [CLASS].`, `a photo of my clean [CLASS].`, `a photo of my new [CLASS].`, and `a photo of my old [CLASS].`. All text anchors are precomputed and cached, but at training and inference time only anchors from classes observed up to the current task are exposed to the model.

**More details.** Following TPPT (Lu et al., 2025), unless otherwise specified, all experiments use the OpenCLIP ViT-B/16 backbone initialized with the `laion400m_e32` checkpoint pretrained on LAION-400M. The complete Stage-I objective uses $\lambda_{\text{align}} = 1.0$, $\lambda_{\text{sup}} = 0.3$, and $\lambda_{\text{neg}} = 0.2$. We set both the text-alignment temperature and the supervised contrastive temperature to

$$\tau_{\text{text}} = \tau_{\text{sup}} = 0.07 \tag{30}$$

for all datasets and tasks. The margin of the negative similarity loss is fixed to $m = 0$, such that only positive similarities to previously observed text anchors are penalized.

At inference, the image and text features are first $\ell_2$-normalized, and the text logits are computed as

$$\boldsymbol{\ell}^{\text{text}} = \gamma_{\text{text}} \mathbf{z} \mathbf{T}_{1:k}^{\top}, \qquad \gamma_{\text{text}} = \frac{1}{0.05} = 20. \tag{31}$$

Here, $\gamma_{\text{text}}$ is a fixed inference temperature scale introduced by our implementation; it is neither inherited from CLIP's learnable logit-scale parameter nor optimized during continual learning. It is also distinct from the training temperature $\tau_{\text{text}} = 0.07$. The auxiliary cosine classifier has its own learnable scale $\gamma_{\text{vis}}$, initialized to 10.

## D Why Incremental LoRA Distillation Is Effective

The goal of consolidation is to approximate explicit multi-module feature composition using a single LoRA. We provide an intuitive explanation from the perspectives of incremental optimization, low-rank adaptation, and empirical convergence diagnostics.

**Incremental residual view.** At task $k$, the teacher target is defined as

$$\mathbf{z}_T^{(k)} = \text{Norm}\left(\frac{k-1}{k}\mathbf{z}_{\text{prev}} + \frac{1}{k}\mathbf{z}_{\text{cur}}\right), \tag{32}$$

where $\mathbf{z}_{\text{prev}}$ is produced by the merged LoRA representing the previous $k-1$ tasks, and $\mathbf{z}_{\text{cur}}$ is produced by the current task-specific LoRA. The student is initialized from the previous merged LoRA and optimized to reproduce this target. Equivalently, before the final normalization, the target can be written as

$$\mathbf{z}_T^{(k)} = \text{Norm}\left(\mathbf{z}_{\text{prev}} + \frac{1}{k}\left(\mathbf{z}_{\text{cur}} - \mathbf{z}_{\text{prev}}\right)\right). \tag{33}$$

Therefore, the student does not need to learn the accumulated representation from scratch. Instead, it fits a task-count-scaled residual relative to the previous merged representation. The coefficient of this residual decreases as more tasks are learned, while historical knowledge is inherited through both the student initialization and the dominant weight assigned to the previous teacher.

**Low-complexity adaptation space.** The effectiveness of this approximation is also related to the restricted adaptation space. In our implementation, all teacher and student branches share the same frozen CLIP backbone and differ only in their visual LoRA parameters. With rank $r = 4$, LoRA is inserted into three projections in each of the 12 visual Transformer blocks, resulting in only $442{,}368$ trainable parameters, approximately $0.30\%$ of the complete model.

The important point is not only that LoRA contains relatively few parameters, but also that both teacher representations are generated by low-rank perturbations of the same frozen backbone. Under a local approximation,

$$\mathbf{z}(\mathbf{x}; \Delta) \approx \mathbf{z}_0(\mathbf{x}) + \mathbf{J}_{\mathbf{x}}\Delta, \tag{34}$$

where $\mathbf{z}_0(\mathbf{x})$ denotes the frozen-backbone representation and $\mathbf{J}_{\mathbf{x}}$ describes the local response to the LoRA parameters $\Delta$. Under this view, the teacher target is constructed from feature perturbations induced within the same low-rank adaptation space. This provides an intuitive explanation for why the target may be easier for another LoRA to approximate than representations produced by two independently fine-tuned full visual encoders.

**Optimization diagnostics.** The optimization logs support this interpretation. On CIFAR-100, the post-consolidation student–teacher cosine similarity ranges from $0.9972$ after the first consolidation to $0.9998$ at later stages. Meanwhile, the final-epoch average distillation objective decreases from $1.7 \times 10^{-3}$ to approximately $1.0 \times 10^{-4}$ across the incremental stages. On ImageNet-R, the recorded cosine similarity remains between $0.9999$ and $1.0000$, and the final loss rounds to $0.0000$ at four-decimal precision for most stages.

Because both student and teacher features are normalized, the feature-matching objective directly constrains their directions, which are the quantities used by CLIP's cosine-similarity classifier. These similarities are diagnostic measurements computed on a mini-batch from the current-task training loader. They therefore show that the student can closely fit the available teacher target on current-task inputs, whereas the preservation of historical knowledge is evaluated separately through the continual-learning results.

**Potential implicit regularization.** Consolidation may also explain why the single merged LoRA occasionally performs slightly better than explicit multi-LoRA composition. Explicit composition retains not only the useful contribution of every task-specific LoRA, but also its task-dependent noise and residual responses to unrelated concepts. When independently adapted features are directly combined, such undesired components may accumulate and introduce interference.

In contrast, consolidation can be interpreted as a data-dependent projection:

$$\Delta^{(k)}_{\text{merge}} = \arg\min_{\Delta \in \mathcal{L}_r} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_k} \left[ d\left( \mathbf{z}(\mathbf{x}; \Delta), \mathbf{z}^{(k)}_T(\mathbf{x}) \right) \right], \tag{35}$$

where $\mathcal{L}_r$ denotes the function family induced by rank-$r$ LoRA modules and $d(\cdot, \cdot)$ denotes the feature-distillation distance. This projection may favor directions that are consistently expressed by the composed teacher on the available inputs, while attenuating variations that cannot be represented consistently by the single low-rank student. Feature distillation can therefore be viewed as ensemble compression with a potential implicit-regularization effect, which may yield a smoother representation than directly combining independently adapted outputs.

This interpretation may help explain why consolidation introduces no clear performance penalty and even yields small improvements on several metrics in Table 3 of the main paper. Compared with explicit multi-LoRA composition, DISCIL improves the final accuracy on CIFAR-100 from $80.57\%$ to $80.86\%$. On ImageNet-R, the average and final accuracies increase from $87.51\%$ and $82.34\%$ to $87.75\%$ and $82.54\%$, respectively. These modest gains should not be interpreted as distillation creating additional task knowledge. Instead, they indicate that compressing the composed representation does not necessarily incur an observable approximation penalty and may occasionally improve generalization by attenuating task-specific noise.

Table 5: **Ablation of the auxiliary visual classifier.** We compare the fixed text-anchor classifier, the prototype-initialized cumulative cosine classifier adopted from MG-CLIP (Huang et al., 2025), and their fusion. "Avg." and "Last" denote the average incremental accuracy and final-stage accuracy, respectively.

| Prediction Branch | Enabled Components | | CIFAR-100 | | ImageNet-R | |
|---|---|---|---|---|---|---|
| | Text | Visual | Avg. | Last | Avg. | Last |
| Text-anchor classifier only | Yes | No | 87.71 | 80.04 | 87.10 | 81.78 |
| Auxiliary visual classifier only | No | Yes | 85.95 | 78.21 | 84.42 | 78.95 |
| **Text–visual fusion** | **Yes** | **Yes** | **88.34** | **80.86** | **87.75** | **82.54** |

Together with the continual-learning results, the near-perfect student–teacher similarities indicate that the proposed consolidation can closely approximate the available composition targets while retaining the practical benefits of explicit multi-LoRA composition. At the same time, it maintains constant consolidation cost and enables task-agnostic inference with a single visual forward pass.

## E Ablation of the Auxiliary Visual Classifier

Following MG-CLIP (Huang et al., 2025), we adopt a prototype-initialized cumulative cosine classifier as an auxiliary visual branch. This component is not the core contribution of DISCIL, but is introduced to complement the fixed text-anchor classifier, particularly when class-name semantics cannot fully capture subtle visual differences. We conduct a simple ablation by evaluating the text-anchor branch alone, the auxiliary visual classifier alone, and their logit-level fusion.

As shown in Table 5, the fixed text-anchor branch alone already provides strong continual recognition performance, confirming that CLIP's stable textual space serves as the primary classification space of DISCIL. The auxiliary visual classifier also achieves competitive standalone performance, but remains weaker than the text-anchor branch. Unlike the globally shared textual anchors, its class weights are learned incrementally from visual features, while the weights of previously learned classes are not revisited. Consequently, the visual branch may provide less stable cross-task calibration when used independently.

Combining the two branches consistently achieves the best results. Compared with the text-anchor classifier alone, text–visual fusion improves Avg. and Last accuracy by $0.63$ and $0.82$ percentage points on CIFAR-100, and by $0.65$ and $0.76$ percentage points on ImageNet-R, respectively. The text-anchor classifier provides stable and transferable semantic references across tasks, whereas the prototype-initialized cosine classifier supplies complementary decision boundaries directly grounded in the visual feature space. These results indicate that the auxiliary classifier, following MG-CLIP (Huang et al., 2025), complements rather than replaces the proposed text-anchor-based representation composition.